Working Paper 94-22
Statistics and Econometrics Series 10
June 1994

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

A SUBSAMPLING METHOD FOR THE COMPUTATION OF MULTIVARIATE ESTIMATORS WITH HIGH BREAKDOWN POINT

Jesús Juan and Francisco J. Prieto*

Abstract:

All known robust location and scale estimators with high breakdown point for multivariate samples are very expensive to compute. In practice, this computation has to be carried out using an approximate subsampling procedure. In this work we describe an alternative subsampling scheme, applicable to both the Stahel-Donoho estimator and the estimator based on the Minimum Volume Ellipsoid, with the property that the number of subsamples required is substantially reduced with respect to the standard subsampling procedures used in both cases. We also discuss some bias and variability properties of the estimator obtained from the proposed subsampling process.

Key Words: Multivariate Analysis, Robust Estimation, Outlier Detection

* Juan: Industrial Engineering School, Universidad Politécnica de Madrid.
Prieto: Dept. of Statistics and Econometrics, Universidad Carlos III de Madrid.

A Subsampling Method for the Computation of Multivariate Estimators with High Breakdown Point

Jesús Juan
Laboratorio de Estadística
Escuela Tec. Sup. Ing. Industriales
Universidad Politécnica de Madrid

Francisco J. Prieto
Dept. Estadística y Econometría
Universidad Carlos III de Madrid

Abstract

All known robust location and scale estimators with high breakdown point for multivariate samples are very expensive to compute. In practice, this computation has to be carried out using an approximate subsampling procedure.

In this work we describe an alternative subsampling scheme, applicable to both the Stahel-Donoho estimator and the estimator based on the Minimum Volume Ellipsoid, with the property that the number of subsamples required is substantially reduced with respect to the standard subsampling procedures used in both cases. We also discuss some bias and variability properties of the estimator obtained from the proposed subsampling process.

1 Introduction

Most classical techniques in multivariate analysis are based on the assumption that the observations follow a normal distribution $N(\mu, \Sigma)$, where μ and Σ denote the location and scale parameters of the distribution, respectively. The maximum-likelihood estimators for these parameters are the sample mean and the sample covariance matrix.

The presence of outliers in the sample can introduce arbitrary modifications in the values of these estimators, and consequently, on the results and conclussions of any multivariate analysis technique based on their values. The identification of outliers and the robust estimation of location and scale parameters are thus different sides of the same problem: solving one of these problems automatically provides a solution for the other one.

A measure of the robustness of an estimator is given by its breakdown point ϵ^* (Hampel, Ronchetti, Rousseeuw and Stahel, 1982). For a given estimator T and a sample of size n,

$$X = \{x_1, x_2, ..., x_n\}, \quad x_i \in \Re^p$$

assumed to be in general position, that is, having no more than p points laying on any hyperplane of dimension $p-1$, the *breakdown point* (for the finite sample size case with replacement, Donoho and Huber, 1983) of the estimator T is defined as

$$\epsilon_n^*(T, X) = \frac{1}{n} \min\{m : \sup_{X_m} T(X_m) < \infty\},$$

where X_m denotes the sample obtained after replacing m observations randomly chosen from X with arbitrary values.

The breakdown point for the sample mean and the sample covariance matrix is $\epsilon_n^* = 0$, that is, it is possible to alter by an arbitrary amount the value of both estimators by modifying just one observation in the sample. As a consequence, it would be of interest to define estimators that are less sensitive to the presence of outliers in the sample, even if that property implies a loss in efficiency.

Another condition that is normally required of location and scale estimators is the property of affine equivariance. A location estimator T is affine equivariant if $T(AX+b) = AT(X)+b$ for any full-rank $p \times p$ matrix A, and any vector $b \in \Re^p$. A scale estimator V is affine equivariant if $V(AX+b) = AV(X)A^T$ for any full-rank matrix A. From the point of view of the identification of outliers, this property implies that no affine transformation will be able to mask the outlying observations.

A significant improvement in the solution of the robust estimation and outlier identification problems came as a consequence of the introduction of the M-estimators (Maronna, 1976). For a sample size of n, these estimators have a breakdown point given by (Tyler, 1990):

$$\epsilon_n^* = \frac{1}{p+1} - \frac{1}{n}.$$

Unfortunately, this value becomes less satisfactory as the dimension of the problem increases. Stahel (1981) and Donoho (1982) proposed the first robust location and scale estimator with high breakdown point for any dimension of the problem (asymptotically equal to 0.5). Later on, Rousseeuw (1985) presented another robust estimator based on the Minimum Volume Ellipsoid, having similar properties.

From a computational point of view, both estimators require a prohibitive amount of time to evaluate, even for small problems. As a consequence, in practice only approximate solutions based on subsampling procedures are computed for both cases. These procedures

aim at obtaining subsamples that do not include any outliers. In this work we present a simple subsampling scheme that guarantees a higher probability of obtaining subsamples having this property, while requiring a reduced computational effort.

In Section 2 of this paper we will briefly describe the two estimators mentioned above. Section 3 will present the subsampling method that we propose, together with its main properties. Finally, in Section 4 we present some conclussions from the properties and behavior of the method.

2 High breakdown point estimators

2.1 The Stahel-Donoho estimator

For a given sample of n observations from $\Re^p$, $X = \{x_1, x_2, ..., x_n\}$, the Stahel-Donoho location and scale estimator $(T_{SD}(X), V_{SD}(X))$ is defined as

$$\begin{aligned} T_{SD}(X) &= \frac{\sum_1^n w_i x_i}{\sum_1^n w_i} \\ V_{SD}(X) &= \frac{\sum_1^n w_i (x_i - T_{SD}(X))(x_i - T_{SD}(X))^T}{\sum_1^n w_i}, \end{aligned} \tag{1}$$

where $w_i = w(r_i)$.

$$r_i = \sup_{d \in S_p} \frac{|d^T x_i - \text{med}_j(d^T x_j)|}{\text{MAD}_j(d^T x_j)}, \tag{2}$$

$S_p = \{d \in R^p : ||d|| = 1\}$ and $w(.)$ denotes a weight function (Hampel, Ronchetti, Rousseeuw and Stahel, 1984).

In this context, r_i provides a measure of how reasonable it is to consider the i-th observation, x_i, as an outlier. If x_i is an outlier, for some unidimensional projection, associated to a direction d, the projected observation $d^T x_i$ will also be an outlier. The median and the median of the absolute deviations (MAD) can be used as robust location and scale estimators for the projections, with breakdown points equal to 0.5. The multivariate robust position and scale estimators are then defined as the weighted sample mean and weighted sample covariance matrix, using weights w_i defined as nonincreasing functions of r_i.

The asymptotic breakdown point for this estimator, that is, the breakdown point as $n \to \infty$, is equal to 0.5. Tyler (1994) has studied the value of the finite sample breakdown point with replacement for both this estimator and a slight modification of the MAD estimator, showing that the estimator $(T_{SD}(X), V_{SD}(X))$ attains the highest breakdown point possible for the class of affine-equivariant estimators (Davies, 1987),

equal to

$$\epsilon_n^* = \frac{[(n-p+1)/2]}{n},$$

where $[a]$ denotes the integer part of a; it is assumed that the sample points X are in general position. Maronna and Yohai (1995) show that (T_{SD}, V_{SD}) has an asymptotic convergence rate of order $1/\sqrt{n}$.

To compute each r_i from (2) we would need to solve a global optimization problem with a nonconvex objective function, having in general a large number of local minimizers. The optimization techniques currently available to solve this problem are too inefficient to be of practical use, even for low dimension problems.

To avoid this difficulty, Stahel proposed to compute an approximation to r_i using the following subsampling procedure: Choose randomly p points from the sample X, and compute a direction orthogonal to the hyperplane defined by the p points, d. Repeat this procedure N times and compute r from (2), replacing S_p with this finite set of directions.

The estimator obtained from this procedure is affine equivariant. Maronna and Yohai (1995) show that the breakdown point of the modified estimator coincides with the value for the estimator computed from the exact procedure under certain conditions. Assume that in a sample X we have replaced a number $m = n\epsilon$ of the original points with arbitrary observations; we will denote the modified sample by X_m. The subsampling method guarantees that the estimator will remain bounded for any X_m if in the process we obtain at least p different subsamples that contain no outliers. If the subsampling procedure is perfectly random, the probability of this condition holding is given by

$$P_0 = 1 - \sum_{k=0}^{p-1} \binom{N}{p} \left((1-\epsilon)^p\right)^k \left((1-(1-\epsilon)^p\right)^{N-k}.$$

We assume the probability of generating the same sample twice is negligible.

Table 1 shows the number of subsamples needed to ensure a probability of success equal to $P_0 = 0.95$, for different contamination levels ϵ and different dimensions of the problem, p. The number of subsamples required is independent of n, and it grows exponentially with the dimension of the problem.

Stahel-Donoho $P_0 = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	9	17	30	58	122
6	17	38	87	223	670
8	28	76	225	780	3365
10	42	143	553	2594	16078
20	225	2414	34936	762520	29233500

Table 1. Stahel algorithm: number of subsamples to attain the breakdown point of the exact algorithm with probability equal to P_0.

2.2 The Minimum Volume Ellipsoid estimator

Rousseeuw (1985) introduced the estimator $(T_R(X), V_R(X))$, based on the Minimum Volume Ellipsoid (MVE), defined as follows: $T_R(X)$ is obtained as the center of the minimum volume ellipsoid containing half the observations, and $V_R(X)$ is the matrix of coefficients of the quadratic form defining the ellipsoid, scaled by a factor to ensure consistency for normal observations. The breakdown point of the MVE estimator is $\epsilon^* = 0.5$ for all p. Subsequently, Davies (1987) defined the class of S-estimators; these estimators include the MVE estimator and have breakdown points that are also independent of the dimension of the data.

In order to compute the minimum volume ellipsoid for a sample X with n observations, it would be necessary to consider all the $\binom{n}{[n/2]+1}$ subsamples of size $[n/2]+1$ in X, and then determine the minimum volume ellipsoid for each one of them. The complexity of the computation of the minimum volume ellipsoid makes this procedure infeasible for problem dimensions larger than two. Furthermore, the growth in the number of ellipsoids to be considered makes the method impractical once n becomes sufficiently large.

An approximate solution (Rousseeuw and Leroy, 1987; Rousseeuw and van Zomeren, 1990) is based on computing a large number of ellipsiods that are not too expensive to generate, and then choosing the one having minimum volume. A subsampling procedure similar to the one described for the Stahel-Donoho estimator can be used to obtain these ellipsoids. This procedure generates N random subsamples of size $p+1$ from X; for each subsample the mean vector $\bar{x}_j$ and the variance matrix V_j are computed, and the ellipsoid defined by $\{x : (x-\bar{x}_j)^T V_j^{-1}(x-\bar{x}_j) \leq 1\}$ is scaled to ensure that it contains $h = [n/2]+1$ observations (if $h = [(n-k+1)/2]$ were used, the breakdown point of the estimator would improve slightly).

The number N of subsamples to be generated can be determined from probabilistic arguments. If the breakdown point of the exact estimator must be achieved, we need to have at least one subsample that contains no outliers. If the number of outliers in X is m and we define $\epsilon = m/n$, the probability of having at least one subsample with this property is given by

$$P_1 = 1 - \left(1 - (1-\epsilon)^{p+1}\right)^N.$$

Table 2 shows the value of N for $P_1 = 0.95$ and different values of the contamination level ϵ and the dimension of the problem p.

MVE $P_1 = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	4	8	17	37	95
6	5	13	35	106	382
8	7	21	73	296	1533
10	8	34	150	825	6134
20	26	324	5362	136560	6282506

Table 2. Rousseeuw's algorithm: number of subsamples to attain the breakdown point of the exact algorithm with probability equal to P_1.

2.3 Additional considerations

Other estimators with high breakdown point have been defined: Rousseeuw (see pg. 303 in Hampel. Ronchetti, Rousseew and Stahel, 1986) proposes a variant of the minimun volume ellipsoid, the minimum covariance matrix determinant estimator (MCD). Davies (1987) suggests some modifications for the MVE estimator, while studying its convergence and breakdown point properties for finite samples. Maronna, Stahel and Yohai (1992) present an affine equivariant estimator based on projections, having also a breakdown point that is independent of the dimension of the data. The algorithm suggested for the computation of this estimator is based on a subsampling scheme that can also be modified to use the subsampling scheme proposed in the following section.

An extensive simulation study conducted by Maronna and Yohai (1994) compares the behavior of the different methods described in this section, concluding that the Stahel-Donoho estimator has the best bias and variability properties; this estimator is also the most efficient one for outlier identification under a range of different structures in the distribution of the outliers.

The subsampling approximations described in the preceding paragraphs have been defined with the goal of replicating the breakdown point properties of the corresponding

exact estimator. Any reasonable approximation to the bias and variability properties of the exact estimators would require a significantly higher number of subsamples. These remarks constitute an additional motivation for the development of subsampling methods that require a reduced number of subsamples, while being able to generate a high proportion of "good" subsamples.

3 Proposed subsampling algorithm

Let ϵ denote the proportion of outliers in the sample X; the probability of a subsample of size p generating a "good" direction for the Stahel-Donoho estimator, that is, the probability of the subsample containing no outliers, is given by $(1-\epsilon)^p$, and for a subsample of size $p+1$ for the MVE estimator the probability is given by $(1-\epsilon)^{p+1}$.

The motivation behind the proposed subsampling scheme is to increase the probability of obtaining "good" subsamples. This goal can be achieved by using the following procedure: construct subsamples of size k, remove from each subsample one observation, and take the remaining $k-1$ observations as the final subsample to construct the desired estimator. The final subsample will be a "better" subsample than the original one if the probability of removing an outlier from the initial sample is sufficently high. We now describe a procedure to remove one observation from the subsample having the property that, if the subsample contains just one outlier, then this outlier will be the observation excluded from the subsample. If this procedure is used, the probability that the final subsample contains no outliers is given by

$$(1-\epsilon)^k + k(1-\epsilon)^{k-1}\epsilon.$$

This probability is a decreasing function of k, and it would be optimal to choose k as small as possible. The actual value of k will also depend on the procedure used to select the observation to be removed from the subsample. An additional condition on the whole procedure is that it should be computationally efficient.

Let $\bar{x}_{(i)}$ and $V_{(i)}$ denote the mean and covariance matrix of the modified subsample, obtained by removing observation x_i from the subsample of size k. If observation x_i were the only outlier in the subsample, its distance to the mean, $d_{(i)}$, defined as

$$d^2_{(i)} = (x_i - \bar{x}_{(i)})^T V^{-1}_{(i)} (x_i - \bar{x}_{(i)}),$$

should be larger than $d_{(j)}$ for any $j \neq i$. If x_i is the only outlier in the subsample, both $\bar{x}_{(i)}$ and $V_{(i)}$ are unbiased estimators, unaffected by the contamination in the sample.

The proposed scheme proceeds by removing the observation having the largest value of $d_{(i)}$. If $\bar{x}$ and V denote the sample mean and the sample covariance matrix for the

whole sample, the Mahalanobis distance for observation i, d_i, given by

$$d_i^2 = (x_i - \bar{x})^T V^{-1} (x_i - \bar{x}_i), \tag{3}$$

and $d_{(i)}^2$ are related by

$$d_{(i)}^2 = \frac{(n-2)n^2}{(n-1)^3} \frac{d_i^2}{1 - nd_i^2/(n-1)^2}.$$

This equality implies that $d_{(i)}^2$ is a monotonically increasing function of d_i^2; the largest value of $d_{(i)}$ will be the one corresponding to the largest distance d_i.

For a sample with exactly one outlier, the most powerful test is the one that removes the observation having the largest Mahalanobis distance, d_i.

To apply this procedure we must have a subsample of size at least equal to $k = p+2$. The algorithm that uses the proposed subsampling method to compute the Stahel-Donoho estimator has the following form:

1. Construct N subsamples of size $p + 2$.
2. Remove from each subsample the observation having the largest Mahalanobis distance.
3. Compute the directions orthogonal to each of the $p + 1$ subsets of p observations that can be formed from the final subsample of size $p + 1$.
4. Compute r_i from (2), replacing S_p with the set of directions obtained in step 3.

If the final subsample contains no outliers, this procedure would compute $p+1$ "good" directions from each subsample. If we generate N subsamples, the probability of having at least one that contains no outliers after removing the "worst" observation is given by

$$P_2 = 1 - \left(1 - (1-\epsilon)^{p+2} - (p+2)(1-\epsilon)^{p+1}\epsilon\right)^N.$$

Table 3 shows the number of subsamples required to have $P_2 = 0.95$ for different contamination levels ϵ and different dimensions of the data p.

Stahel-Donoho $P_2 = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	2	3	6	12	26
6	2	5	11	27	84
8	3	7	19	64	278
10	3	10	34	152	943
20	8	61	734	14527	546304

Table 3. Proposed method: number of subsamples to attain the breakdown point of the exact algorithm with probability equal to P_2.

The reduction in the number of subsamples with respect to the values shown in Table 1 is significant. Table 4 indicates the reduction factor in the number of subsamples for the proposed method, that is, the ratio between the number of subsamples required by the traditional approach, as shown in Table 1, and the number of subsamples required by the proposed approach. The computations required to determine the $p+1$ directions for each subsample in the proposed method are naturally more expensive than the computations required by the traditional method, but even if this factor is taken into account (see the Appendix), the proposed method is still significantly more efficient than the traditional subsampling algorithm.

Reduction factor Stahel-Donoho. $P = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	4.2	5.5	4.9	4.8	4.7
6	8.2	7.5	7.9	8.2	8.0
8	9.2	10.8	11.8	12.2	12.1
10	13.8	14.2	16.2	17.0	17.0
20	28.0	39.5	48.0	52.5	53.5

Table 4. Number of subsamples required by the Stahel subsampling algorithm for each subsample needed by the proposed method.

Another important advantage of the proposed algorithm is that the average number of "good" directions is also greatly increased, a result that suggests that the estimator obtained after applying the proposed scheme should have better properties than the traditional one. Table 5 compares the expected number of "good" directions for both methods when $\epsilon = 0.5$ and the number of subsamples taken for each method are the ones given in Tables 1 and 3, respectively. For values of P larger than 0.95 the comparison results are even more favorable to the proposed algorithm.

	Stahel	Proposed
4	8	14
6	10	21
8	13	27
10	16	33
20	28	63

Table 5. Expected number of "good" directions when $\epsilon = 0.5$, for Stahel's method and the proposed algorithm.

This scheme can also be applied to the MVE estimator, in the following manner: obtain subsamples of size $p + 2$, remove the observation with the largest Mahalanobis

distance and compute the elemental ellipsoid corresponding to the remaining $p+1$ observations. The number of subsamples that are needed to ensure with probability 0.95 that at least one of them contains no outliers coincide with the values shown in Table 3. Table 6 shows the ratio of the number of subsamples required by the Rouseeuw and van Zomeren (1990) method, as indicated in Table 2, to the number of subsamples required by the proposed method. The computational cost for both procedures is very similar (see the Appendix), implying that the gain in computational efficiency when using the proposed algorithm is even more significant than in the case of Stahel's method.

Reduction factor MVE. $P = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	2.0	2.6	2.8	3.1	3.8
6	2.5	2.6	3.2	3.9	4.5
8	2.3	3.0	3.8	4.6	5.5
10	2.6	3.4	4.4	5.4	6.5
20	3.3	5.3	7.3	9.4	11.5

Table 6. Number of subsamples required by the Rousseeuw and Van Zomeren subsampling algorithm for each subsample needed by the proposed method.

For the number of subsamples indicated in Tables 2 and 6, the expected number of ellipsoids obtained from subsamples with no outliers is similar for both methods and very small (≈ 3). This fact may explain the high bias and variability of the MVE estimator, as mentioned in Cook and Hawkins (1990), Maronna, Stahel and Yohai (1992) and Maronna and Yohai (1995). The proposed subsampling method could be very effective in this sense, as for a given computational cost the expected number of "good" ellipsiods would be increased in the proportion shown in Table 6.

Simulations

When the procedure described in this section is applied to the computation of the Stahel-Donoho estimator, it generates $p+1$ directions for each subsample. Each direction is obtained from $p+1$ points, and any pair of directions from a given subsample shares p common points, implying a certain "dependence" structure between the directions. Although the breakdown point is not affected by this fact, it might have some influence on other properties of the estimator, such as its bias or variability.

To analyze the influence of this "dependence" between directions we have conducted a limited simulation study, comparing both subsampling schemes. For a given normal distribution with parameters μ and Σ (this study can be easily extended to any ellipsoidal

model) we analyze the effect of an ϵ-contamination, generated from an arbitrary distribution G, on the estimators (T_{SD}, V_{SD}). Maronna and Yohay (1994) define as a measure of the bias in the position estimator, $\text{bias}(T_{SD}, G) = (T_{SD} - \mu)^T \Sigma^{-1} (T_{SD} - \mu)$, and for the variance estimator V_{SD}, $\text{bias}(V_{SD}, G) = \varphi(L V_{SD} L^T)$, where φ denotes some measure of nonsphericity and $L^T L = \Sigma^{-1}$ (the Choleski factor of Σ^{-1}). The most common measure of nonsphericity for a matrix A is the condition number $\text{cond}(A)$, defined as the square root of the ratio between the largest and smallest singular values of A. Another measure, used in this simulation study, is

$$\varphi_0(A) = \frac{(\text{tr}(A)/p)^p}{\det(A)},$$

that is, the ratio between the arithmetic and geometric means of the eigenvalues of A, raised to the p-th power. The lower bound for φ_0 is 1, corresponding to the case in which all eigenvalues are equal (sphericity).

Following Maronna and Yohai (1994) we have chosen:

- The most unfavourable contamination model (all outlier observations are concentrated in one point); a sample of n observations with $n - m$ observations taken from an $N_p(0, I)$ distribution (the affine equivariance property of the estimator implies no lack of generality in taking $\mu = 0$ and $\Sigma = I$), and m observations concentrated in $k e_1$, with $m = [n\epsilon]$ and $e_1^T = (1\ 0 \ldots 0)$.
- The Huber function minimizing maximum bias,

$$w(r) = I_{\{r \le c\}} + \frac{c^2}{r^2} I_{\{r > c\}},$$

  where $c = \sqrt{\chi_p^2(0.95)}$, as the weight function in (1).

Figure 1 shows the Box plot corresponding to the results obtained for $p = 6$, $n = 30$, $\epsilon = 0.2$ and $k = 50$. Other values of p, n, ϵ and k give similar results, and this seems to indicate that there is no significant loss in the "quality" of the directions generated by the proposed subsampling method, due to the close relationship that exists between the directions obtained from a given subsample.

4 Conclusions

Several robust estimators for the position and scale parameters of a multivariate normal sample, with good theoretical properties regarding convergence, efficiency, bias and breakdown point for highly contaminated samples, have been proposed in the literature. None

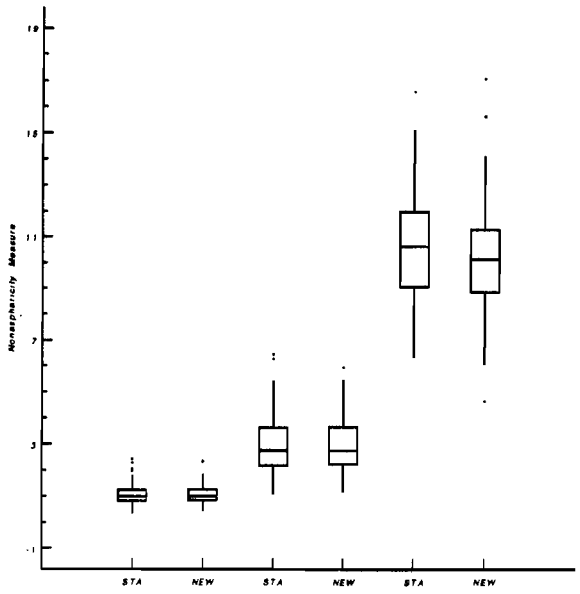

Figure 1: Log of sphericity measure for the standard and proposed subsampling schemes.

of these estimators can be computed in exactly the form they have been defined, and all of them must be approximated by procedures based on subsampling schemes. In this paper we have presented a new subsampling procedure that requires a much smaller number of subsamples. By taking advantage of this property, it would be possible to obtain a much better estimator at a lower computational cost. The estimators obtained in this manner are able to detect complex contamination patterns in the sample.

References

[1] Atkinson. A.C. and Mulira, H.-C. (1993), " The stalactite plot for the detection of multivariate outliers," *Statistics and Computing*, 3, 27-35.

[2] Davies, P.L. (1987), "Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269-1292.

[3] Donoho, D.L. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph. D. Qualifying Paper, Harvard University.

[4] Golub, G.H. and Van Loan, C.F. (1989) *Matrix Computations*, Baltimore: The Johns Hopkins University Press.

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Appoach Based on Influence Functions*, New York: John Wiley and Sons.

[6] Maronna, R.A., Stahel, W.A. and Yohai, V.J. (1992), "Bias-Robust Estimators of Multivariate Scatter Based on Projections," *Journal of Multivariate Analysis*, 42, 141-161.

[7] Maronna, R.A. and Yohai, V.J. (1994), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," To appear in the *Journal of the American Statistical Association.*

[8] Rousseeuw, P.J. (1985) " Multivariate Estimators With High Breakdown Point," in *Mathematical Statistics and its Applications* (vol. B), eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Dordrechts-Reidel, 283-297.

[9] Rousseeuw, P. and van Zomeren, B.C. (1990), " Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-651.

[10] Rousseeuw, P. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley and Sons.

[11] Stahel, W.A. (1981), "Breakdown of Covariance Estimators," Research Report 31, Fachgruppe fur Statistik, E.T.H. Zurich.

[12] Tyler, D.E. (1987), "A distribution-free M-estimator of Multivariate Scatter," *The Annals of Statistics.* 15, 234-251.

[13] Tyler, D.E. (1994), "Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics," To appear in *The Annals of Statistics.*

Appendix. Evaluation of computational costs

In Section 3 it was mentioned that the computational costs of the different subsampling schemes should be taken into account when comparing the performance of the procedures. For example, it is important to consider this computational cost when analyzing the results shown in Tables 4, 5 or 6. In this Appendix we evaluate these computational costs for both the Stahel-Donoho estimator and the MVE estimator, and we include these costs into the analysis of the results presented in the paper.

A detailed evaluation should take into account the hardware to be used and details of the implementation of the algorithm, for example; as we are interested only on approximate measures of efficiency, we will only consider in this appendix an estimate of the numbers of arithmetic operations (sums and products) required for efficient implementations of the different methods, ignoring the cost of control instructions, comparisons, ... The numbers of operations for basic numerical procedures can be obtained from standard references on numerical linear algebra (Golub and Van Loan, 1989).

We will assume throughout that we have been given a sample of size n in a space of dimension p.

The Stahel-Donoho estimator

Proposed procedure

The subsampling procedure proposed in the paper would obtain the estimator from the following steps:

1. Select a subsample of $p+2$ observations.
2. Compute the subsample mean $\bar{x}$ and covariance matrix V.
3. Compute the Mahalanobis distance for each observation in the subsample using (3). We first compute the Choleski factor of the covariance matrix V, R, then solve the system $Ru_i = x_i - \bar{x}$, and finally form $u_i^T u_i$.
4. Remove from the subsample the observation with the largest Mahalanobis distance.
5. Compute the projections of all points in the sample along the directions orthogonal to each subset of p points from the subsample, d_l, $l = 1, \ldots, p+1$.

   Let W_{jk} denote the matrix whose rows are the vectors $x_i - x_k$ for some observation k in subsample j and all observations $i \neq k$. The orthogonal direction d_l, $l = 1, \ldots, p$,

can be obtained as the solution of the system of equations $W_{jk}d_l = e_l$, where e_l is the l-th unit vector. We can compute p orthogonal directions as the columns of the matrix D_j solution of the system of equations $W_{jk}D_j = I$. The projections of sample point x_i along these p directions corresponding to subsample j can be obtained as the components of the solution of the system of equations $W_{jk}^T q_{ji} = x_i$. The $p+1$-st orthogonal direction is given by $d_k = -\sum_j d_j$, and the corresponding projection can be obtained as $-e^T q_{ji}$. Note that only one observation in the subsample needs to have its projection computed.

6. For each set of projections, compute the median and the MAD, and form the weights r_i from (2).

7. Finally, obtain the values of $(T_{SD}(X), V_{SD}(X))$ from (1).

The following table summarizes the costs of these steps:

Step	Operation	Cost
2	$x_i - \bar{x}$	$2p(p+2)$
	Covariance matrix	$(p+2)(p+1)p$
3	Choleski factorization	$p^3/3$
	Computation of u_i	$(p+2)p^2$
	Computation of $\|u_i\|^2$	$2(p+2)p$
5	LU factorization of W_{jk}	$2p^3/3$
	Solution of $W_{jk}^T q_{ji} = x_i$	$2(p^2-p)(n-p)$
	$p+1$-st projection	$p(n-p)$
6	Computation of r_i	$2n$
7	$T_{SD}(X)$	$2np+n$
	$V_{SD}(X)$	$np(p+1)+2np$

The total cost is given by

$$N_1(2np^2 - np + 2n + p^3 + 10p^2 + 8p) + np^2 + 5np + n,$$

where N_1 denotes the number of subsamples generated by the algorithm.

Stahel's procedure

This procedure is similar to the one described above, except that now the subsample has only p observations, steps 2, 3 and 4 are not needed, and step 5 is replaced by

5. Compute the direction orthogonal to all pairs of observations in the subsample.

   As in the proposed algorithm, let W_{jk} denote the matrix whose rows are the vectors $x_i - x_k$ for some observation k and all observations $i \neq k$ in subsample j. The orthogonal direction d_j can be obtained as a non-zero solution for the system of equations $W_{jk}d_j = 0$, computed from an LU factorization of W_{jk}. Obtain the projections of all sample points onto this direction, $d_j^T x_i$.

The costs of these steps are

Step	Operation	Cost
5	LU factorization of W_{jk}	$p(p-1)^2 - (p-1)^3/3$
	Computation of d_j	$2(p-1)^2 - (p-1)$
	Computation of $d_j^T x_i$	$2(n-p+1)p$
6	Computation of r_i	$2n$
7	$T_{SD}(X)$	$2np + n$
	$V_{SD}(X)$	$np(p+1) + 2np$

If N_2 denotes the total number of subsamples, the number of operations for all steps will be approximately equal to

$$N_2(2np + 2n + \frac{2}{3}p^3 - p^2 - 3p) + np^2 + 5np + n.$$

In Table 7 we show the ratio of the computational cost required by the Stahel subsampling method and the computational cost of the proposed scheme when both procedures generate the number of subsamples needed to guarantee the breakdown point of the Stahel-Donoho method with probability 0.95, as shown in Tables 1 and 3. We assume that $n = 5p$ in all cases.

Ratio computational cost Stahel-Donoho. $P = 0.95$					
$p\backslash\epsilon$	0.1	0.2	0.3	0.4	0.5
4	1.1	1.3	1.2	1.2	1.2
6	1.5	1.4	1.5	1.6	1.5
8	1.4	1.7	1.9	1.9	1.9
10	1.8	1.9	2.3	2.4	2.4
20	2.7	3.9	4.8	5.3	5.4

Table 7. Ratio of operations required by the Stahel subsampling algorithm and the proposed method.

We could also compare the expected number of "good" directions that can be obtained for both methods for the same computational cost. Assume that we compute the number of subsamples given in Table 1 for the Stahel procedure, and that for the proposed algorithm we generate a number of subsamples such that the computational cost is the same. Table 8 gives the average number of good directions generated by Stahel's method and the proposed algorithm for that fixed computational cost.

	Stahel	Proposed
4	8	17
6	10	31
8	13	52
10	16	79
20	28	338

Table 8. Expected number of subsamples with no outliers when $\epsilon = 0.5$, for Stahel's method and the proposed algorithm. Equal computational effort.

The MVE estimator

Proposed procedure

The proposed subsampling procedure would have to perform the following operations:

1. Select a subsample of $p + 2$ observations.
2. Compute the subsample mean $\bar{x}$ and covariance matrix V.
3. Compute the Mahalanobis distance for each observation in the subsample using (3). Use the Choleski factor of V.
4. Remove from the subsample the observation with the largest Mahalanobis distance.
5. Compute the mean and covariance matrix for the modified subsample. Update the Choleski factor.
6. Compute the value of d_i^2, using (3) with $\bar{x}$ and V the values for the subsample, for all points in the sample, and obtain the median of these values d_m.
7. Compute the volume of the ellipsoid from d_m and the determinant of V, from its Choleski factor.

8. Finally, obtain the values of $(T_R(X), V_R(X))$ from the ellipsoid having minimum volume from all the ones generated in the subsamples.

The following table summarizes the costs of these steps:

Step	Operation	Cost
2	$x_i - \bar{x}$	$2p(p+2)$
	Covariance matrix	$(p+2)(p+1)p$
3	Choleski factorization	$p^3/3$
	Computation of u_i	$(p+2)p^2$
	Computation of $\|u_i\|^2$	$2(p+2)p$
5	Update $\bar{x}$	$2p$
	Update Choleski factor	$5p^2$
6	Computation of d_i^2	$(n-p-1)(p^2+3p)$
7	Computation of $\det(V)$	p

If N_3 denotes the number of subsamples considered, the total number of operations for all steps will be approximately equal to

$$N_3(np^2 + 3np + \frac{4}{3}p^3 + 10p^2 + 10p).$$

Rousseeuw and Van Zomeren procedure

This procedure is very similar to the preceding one, except that now we only have $p+1$ points in the subsample, and steps 2, 3 and 4 are no longer needed.

If N_4 denotes the number of subsamples to be taken, after removing the cost of steps 2, 3 and 4 from the preceding total we obtain

$$N_4(np^2 + 3np + \frac{1}{3}p^3 + 3p).$$

Finally, the following table shows the comparison of computational effort required by the Rousseeuw and van Zomeren method and the proposed algorithm, for $P = 0.95$.

Ratio computational cost MVE. $P = 0.95$					
$p \backslash \epsilon$	0.1	0.2	0.3	0.4	0.5
4	1.4	1.9	2.0	2.2	2.6
6	1.8	1.9	2.3	2.9	3.3
8	1.8	2.3	2.9	3.5	4.2
10	2.0	2.6	3.4	4.2	5.0
20	2.6	4.2	5.8	7.5	9.2

Table 9. Ratio of operations required by the Rousseeuw and van Zomeren subsampling algorithm and the proposed method.